# Thermodynamics and Statistical Mechanics

*Nic Ford*

## 1 Introduction

This article is part of a series on physics for mathematicians. It's about the physics of macroscopic systems, objects on the scale that you might interact with in everyday life. While the behavior of macroscopic objects should, in principle, be completely explainable in terms of their microscopic components, it's often far from clear how this is supposed to work. What exactly does a quantity like temperature correspond to on the microscopic level? How do we account for the fact that many macroscopic phenomena seem to happen in only one direction, while the microscopic physics is completely time-reversible?

There are two closely related areas of physics that touch on these questions: *thermodynamics* is the high-level description of macroscopic physics, and *statistical mechanics* is the framework by which we can extract this description from the underlying microscopic laws. This is the part of physics that has the most to say about the sorts of physical objects human beings ordinarily interact with, and given how large and complicated these objects are, it's surprising how well it can be understood.

Thermodynamics and statistical mechanics might be further divided into the *equilibrium* and *non-equilibrium* theories. The equilibrium theories are concerned with physical systems that have reached the point where their macroscopic properties are not changing over time, whereas the non-equilibrium theories describe moments when these macroscopic properties are still changing. (In particular, *how* a system gets to equilibrium in the first place is a question for non-equilibrium statistical mechanics.) We focus mostly on the equilibrium case, which is much better behaved theoretically, with just a few qualitative comments on the non-equilibrium case. I may cover the non-equilibrium theory in more detail in a future piece.

This article is sort of an odd fit for the series. The mathematics involved is less complicated than the other articles in the series, but I still found this subject quite difficult to learn. Perhaps because it's so grounded in the "everyday world," it doesn't lend itself to the sort of crisp presentation mathematicians tend to like, with everything following from a short list of axioms. I have reluctantly concluded that a strictly axiomatic approach would be more confusing than helpful, so, while I have still tried to make everything feel natural, there are many points where some input from the physical world is required to make sense of things.

This difficulty is compounded by the fact that there isn't really a consensus on the "correct" foundations for statistical mechanics. There is a quote from the article "A Field Guide to Recent Work on the Foundations of Statistical Mechanics" by Roman Frigg that I think sums up the situation well:

> Unlike quantum mechanics and relativity theory, say, SM [statistical mechanics] has not yet found a generally accepted theoretical framework, let alone a canonical formulation. What we find in SM is a plethora of different approaches and schools, each with its own programme and mathematical apparatus, none of which has a legitimate claim to be more fundamental than its competitors.

I have made a choice of theoretical framework which seems well-motivated mathematically, but there's no reason to take that choice as an argument in favor of some philosophical position; there definitely are alternatives. That same survey article is a good overview of the options.

A final difficulty is that, while a lot of work has been done on building a complete, mathematically rigorous version of statistical mechanics, this work is not at all complete. I have indicated what can be done rigorously to the best of my knowledge, but along the path from microphysics to thermodynamics we will sometimes have to make the logical leap of just assuming that some step works out the way we'd want it to. I've written a companion piece to this article in which I analyze a very simple toy model in which the whole process can be done rigorously from start to finish, which might at least help you see how the picture is supposed to look for more realistic systems.

With all that said, even the equilibrium version of statistical mechanics that we develop here is shockingly useful; perhaps because it assumes so little about the details of the microphysics, the core ideas can be applied to a very large number of situations. In addition, an eventual goal of this series of articles is to build up to a presentation of quantum field theory, and many pieces of the quantum field theory story show up in a somewhat simpler form in statistical mechanics, and so it's worth getting a handle on it for that reason as well.

I found the following books and articles helpful when preparing this piece:

- *An Introduction to Thermal Physics* by Daniel V. Schroeder is an undergraduate physics textbook, and it therefore doesn't use any math more complicated than a partial derivative. I found it to be a very good source of physical intuition, and I'd recommend it for that reason; it should be a pretty easy read for anyone who has been following this series.
- *Mathematical Statistical Mechanics* by Colin J. Thompson is a book from the 1970's pitched at about the same level as this article. Unfortunately it was a bit difficult for me to find a copy, but it is worth a read.
- Edwin Jaynes was a physicist who advocated a point of view on statistical mechanics that I would call radically Bayesian. I don't completely align with him philosophically, but his perspective still influenced this article quite a bit, and I also just found him enjoyable to read. I recommend the short article "Information Theory and Statistical Mechanics" and the longer set of lecture notes "Where Do We Stand on Maximum Entropy?".
- Large deviation theory offers a useful perspective (that we won't touch on here at all) on the equilibrium distributions we will discuss in the second half of the article, and I found two articles by Hugo Touchette useful for learning about it: "The large deviation approach to statistical mechanics" and "Equivalence and nonequivalence of ensembles".
- *Modern Thermodynamics* by John Denker is a somewhat loosely organized free book that contains a lot of intuition that I found helpful.
- The survey article I quoted above is "A Field Guide to Recent Work on the Foundations of Statistical Mechanics" by Roman Frigg. It's a good place to get a sense of where philosophers stand on some of the foundational questions that I only gesture at briefly in this article.

I am very grateful to Jordan Watkins and Yuval Wigderson for many helpful suggestions on earlier versions of this article.

# 2 Thermodynamics

Our ultimate goal is to describe how macroscopic phenomena like temperature and pressure arise from the very different-looking microscopic description of physics that is supposed to underlie it. Before we do this, though, I think it's helpful to have a firm idea of what this macroscopic picture actually looks like so we can know what it is we're aiming for.

This section is an introduction to *thermodynamics*, the name for this macroscopic description of the physics of temperature, pressure, heat, work, and so on. We will work entirely at the macroscopic level, with no reference to microphysics and with the irreversibility of the dynamics "baked in" from the start.

Again, no attempt is made to give a strict axiomatic presentation. This is an emergent, high-level theory and we'll have to refer to actual physical objects a fair amount. If you would like to see what a more rigidly axiomatic version of this story might look like, you can read "The Physics and Mathematics of the Second Law of Thermodynamics" by Lieb and Yngvason. (They also have a shorter version of the piece called "A Guide to Entropy and the Second Law of Thermodynamics".)

## 2.1 Equilibrium and Thermodynamic States

Thermodynamics is a description of physics on the *macroscopic* level, and as such it's completely agnostic about anything having to do with the fundamental constituents of matter. (In fact, much of the theory was developed at a time when the idea that matter is made of atoms was still controversial!) The basic object of study is a **system**, which can be taken to refer to basically any macroscopic object, from a box of gas on a table to a steam engine to the earth's atmosphere. We will often distinguish between a **composite stystem**, which can be divided into **subsystems**, and a **simple system** which cannot. The decision of whether or how to divide a system into subsystems depends on the problem you are trying to solve; you might, for instance, choose to divide a gas in a large container into small cubes and track the properties of each piece separately, or just consider the gas as an indivisible whole.

Probably the central idea of thermodynamics (at least the way I am presenting it) is **equilibrium**. Physically, a system is in equilibrium when the values of the relevant measurable quantities have mostly stopped changing on the time scale you are interested in. At this level of abstraction, equilibrium should be thought of as one of the fundamental concepts in the theory, rather than as something expressible in terms of simpler notions.

Questions like how equilibration happens on the microscopic level, what time scale is relevant, or how large a fluctuation can be before the system hasn't "mostly stopped changing" are not ones that thermodynamics answers. Instead, we will just assume that, given enough time, every system will eventually come to equilibrium. Good examples of equilibration to keep in mind are hot soup cooling down to room temperature on a table; or a gas, initially confined to one half of a box, expanding to fill the whole box uniformly.

The state of a system in equilbrium can be specified by listing the values of some small number of **thermodynamic variables**. These variables are quantities like total energy, temperature, pressure, volume, angular momentum, number of particles, and so on; the exact list of relevant

thermodynamic variables depends on the system under consideration. The possible equilibrium states of a given system correspond to points on a manifold called the **thermodynamic state space**, which for us will always be some $\mathbb{R}^n$; the thermodynamic variables are then just real-valued functions on the state space. Thermodynamic variables always correspond to quantities that can be measured in an experiment that could actually be practically performed. A quantity like "the pressure the gas exerts on the left wall of the box" is a suitable thermodynamic variable; "the velocity in meters per second of the gas particle that was closest to the top of the box at 10:00 this morning" is not.

It is best to think of this setup — in which the system is assumed to reach a unique equilibrium state characterized by a small number of thermodynamic variables — not as an assertion about how the whole world works but as a rule for determining which systems we intend to analyze with the tools of thermodynamics at all. We can say that a "thermodynamic system" is something that behaves in this way; there certainly are non-thermodynamic systems in the world, and thermodynamics is not a good description of them! The claim that equilibration always happens is sometimes somewhat playfully called the "minus-first law of thermodynamics."

In a composite system, we might speak of the value of some thermodynamic variable for one subsystem or another. For example, in a system consisting of a hot bowl of soup together with the cooler air around it, we can ask about the temperature of the air or the temperature of the soup. If the soup and the air are allowed to interact in the way they would in the real world, this composite system is not in equilibrium. (As we will see soon, at equilibrium they have the same temperature.) Once the composite system has equilibrated we will often say that the soup is **at equilibrium with** the air. The **zeroth law of thermodynamics** is the assertion that this is an equivalence relation; transitivity is really the only nontrivial claim here.

Note that it is therefore a slight violation of our rules to speak about the temperature of the hot soup, since thermodynamic variables only have well-defined values for systems in equilibrium! Nonetheless, this rule-breaking is completely pervasive, and is in fact necessary to do much of anything interesting with the theory. In this situation, you should imagine that, while energy is flowing between the soup and the air, this happens much more slowly than it takes for the soup to equilibrate on its own, so at any moment in time we may pretend that the soup is at equilibrium. Because it's so much easier to describe equilibrium states than non-equilibrium states, simplifying assumptions of this type will come up a lot.

The exact form of the state space — in other words, the answer to the question of which variables suffice to describe the state of a system in equilibrium — is outside the purview of thermodynamics itself. By writing down a complete list of thermodynamic variables for a system we are *asserting* that this list contains enough variables to predict the future behavior of the system for whatever purposes we're interested in. Once we have such a list, the laws of thermodynamics give us constraints on how the values of the variables can change, but they don't tell us which variables to use ahead of time.

It will often happen that the values of some thermodynamic variables will be completely determined by the others. Such a relationship is called an **equation of state**. One famous example of an equation of state is the **ideal gas law**, written $PV = NkT$, which holds for a gas in equilibrium which is sparse enough that interactions between the gas particles can be neglected. Here $P$ is pressure, $V$ is volume, $N$ is the number of gas particles, $T$ is the temperature, and $k$ is **Boltzmann's constant**. (Boltzmann's constant is approximately $1.38 \times 10^{-23}$ J/K; we will see in the statistical mechanics section that it plays a fundamental role in the theory.) Like the complete list of thermodynamic variables, any equations of state are an *input* to thermodynamics, not a prediction. Many equations of state can be derived using the machinery of statistical mechanics, and in fact we will do this for the ideal gas law at the end of this article.

In general, there should be no expectation that the same list of variables that suffices to pick out an equilibrium state will determine everything interesting about a non-equilibrium state. For example, while a box of gas at equilibrium has a single temperature, if the temperature is *not* uniform then the particular temperature distribution can certainly have macroscopically noticeable effects; and to describe the state of a container of water close to freezing, it's probably necessary to know something about the relative amounts of ice and liquid water at any moment in time.

For this and other reasons, non-equilibrium thermodynamics is more challenging to describe theoretically, and (in my opinion) addressing some of these issues comes at some cost to elegance. So, again, we'll mostly stick to equilibrium thermodynamics in this article. The prototypical equilibrium thermodynamics question is something like the following. Suppose a system starts in an equilibrium state with some known values of the thermodynamic variables, but then we change the constraints and allow it to come to equilibrium again. What are the resulting values of the variables?

## 2.2 Energy and Entropy

While the exact list of thermodynamic variables depends on the problem, there are two that will always show up. The first is **energy**, denoted by $E$. This is the same quantity that is referred to as "energy" in Newtonian mechanics. In particular, it is always conserved, which is usually the main reason it's interesting to keep track of. Within the context of thermodynamics, the conservation of energy is sometimes called the **first law of thermodynamics**, although this name is sometimes instead attached to a *consequence* of energy conservation that we will see in just a moment.

(If you learn more about thermodynamics, you may encounter a quantity called "internal energy," denoted by $U$. This refers to the energy "contained within the system," excluding the kinetic and potential energy associated with the motion of the system's center of mass. I don't find this distinction very helpful, especially for the ideas we will consider in this article. When a clear distinction can be made between $U$ and $E$, one can often just keep track of $E$ and just say specifically which forms of energy are relevant for which purposes.)

In a composite system, it can be useful to talk about the energy of one subsystem or another, and it is common to assume that the energy of the whole system is the sum of the energies of its components. It's important to realize that this is an approximation; in fact, unless the components are completely isolated from each other, it is not possible to divide all of the energy into subsystems in this way. Think of two bodies interacting via Newtonian gravity. The total energy is the sum of three terms: the kinetic energy of the first body; the kinetic energy of the second body; and the gravitational potential energy, which depends on the locations of *both* bodies and so can't be assigned to just one of them.

A common case in which this approximation is appropriate is when the subsystems are in **thermal contact** with one another. This means that they are able to exchange energy, but that the energy associated with their interaction is very small compared to the energies of each system separately and so can be neglected. A good example is a box of gas in contact with the air. The energy of the gas grows with the volume of the box, but the energy associated with the interaction grows with the surface area, and so for an appreciably-sized box it will be much smaller.

The second important thermodynamic variable is called **entropy**, denoted by $S$ and usually measured in units of energy divided by temperature, like J/K. The most important thing about entropy is the famous **second law of thermodynamics**. For our purposes, it says that for an

isolated system, if we start in an equilibrium state, then change the constraints of the system somehow and allow it to reach equilibrium again, the entropy of the final state is greater than or equal to the entropy of the initial state.

Once we've defined temperature and heat we will talk about how one might actually measure entropy in practice. Later, in the section on statistical mechanics we will talk about what entropy "actually is," but at this level of abstraction it is just a thermodynamic variable to which the second law applies. It might help, in fact, to temporarily set aside any ideas you may have had, especially any having to do with it being a "measure of disorder," until we can address the question properly. (This perspective also has the advantage of being truer to the history — thermodynamic entropy and the second law of thermodynamics predate any interpretation of entropy in terms of statistics!)

Just like with energy, it is common to write the entropy of a composite system as the sum of the entropies of each of its subsystems, and this is again just an approximation which is suitable when the contact between the subsystems is light. As with energy, this is exactly true only when the systems are completely isolated from each other. Because we lack a microscopic definition of entropy at this point, the additivity of entropy will just have to be postulated. (There are also some exotic situations in which it's not even *approximately* true, but for this article we will assume that this never happens.)

This version of the second law — which only compares the entropies of two equilibrium states and says nothing about what's happening in the middle — might seem weaker than you were expecting. Strictly speaking, though, saying anything stronger would require talking about the thermodynamic state of a system that is out of equilibrium, and as we've discussed, this is a much harder problem; defining thermodynamic entropy in the non-equilibrium case is, depending on the assumptions one is willing to make, somewhere between difficult and impossible.

However, it is sometimes necessary to take just one step into the non-equilibrium regime when considering two systems in thermal contact. In this case, we assume that the time it takes to exchange an appreciable amount of energy is much longer than the time it takes for each system to equilibrate, so that we are justified in modelling the two systems as always being separately in equilibrium, just with slowly changing values of the total energy.

In this situation, we have a slightly stronger version of the second law: the equilibrium state of the composite system *maximizes* the sum of the entropies of the two systems separately. (A less formal but often helpful way to think about this is as the claim that any state transition which is compatible with the constraints of the problem and increases the entropy will happen eventually.) This form of the second law will be helpful when we discuss temperature in just a moment.

Suppose we take a system and increase its size by a factor of $m$. Many thermodynamic variables can be usefully placed into one of two categories based on how they behave in this situation. We say a quantity is **intensive** if it stays the same under this rescaling, and **extensive** if it also multiplies by a factor of $m$. Extensive quantities include mass, volume, the number of particles, energy, and entropy; intensive quantities include density, temperature, and pressure.

Most extensive quantities also add when forming a composite system; this is true of everything on this list, and in particular we have already assumed that it is true of entropy. In the case of entropy, it's important that we are asserting this additivity *before* the combined system equilibrates; afterwards the entropy might be higher than the sum of the entropies of the original component systems.

This can be leveraged to demonstrate a useful property of the entropy. Suppose we have parameterized the state space using only extensive quantities, not including the entropy, and

suppose that they are all conserved quantities. (For example, for an ideal gas, we might use energy and volume.) We may then think of $S$ as a real-valued function on the resulting copy of $\mathbb{R}^n$. Consider two points $x$ and $y$ representing two different systems, and let the notation $mx$ denote the result of rescaling $x$ by a factor of $m$. Then for any $m \in [0,1]$, the composite system consisting of $mx$ and $(1-m)y$ has entropy $S(mx) + S((1-m)y) = mS(x) + (1-m)S(y)$, since entropy is extensive. After equilibrating, since the values of the coordinates are conserved by assumption, we are at the point $mx + (1-m)y$. The entropy can't have decreased, so we conclude that

$$S(mx + (1-m)y) \geq mS(x) + (1-m)S(y),$$

that is, under our assumptions, the entropy is *concave*.

## 2.3 Temperature and Pressure

Many other thermodynamic variables — most notably the temperature — can be derived from the energy and entropy. Suppose we find ourselves at an equilibrium state described by some point $x$ in the state space. Consider a system of coordinates around $x$ consisting of the entropy $S$ together with some number of additional thermodynamic variables $V_1, \ldots, V_n$, not including the energy. We will assume that the $V_i$'s are all quantities that are easily measureable macroscopically, like volume. A good example to keep in mind is an ideal gas in a box with a fixed number of particles, for which $n = 1$ and our list of variables consists of just the entropy and the volume $V$. We will assume that the energy is not one of the $V_i$'s.

Imagine then that the conditions change in some way — for example, a small amount of energy is added to the gas by placing it over a flame for a short time — knocking the system slightly out of equilibrium in such a way that, when it equilibrates again, we find ourselves at a new point in state space very close to $x$. We can express the difference in the energy of these two equilibrium states in terms of the differences in the other variables:

$$dE = \frac{\partial E}{\partial S} dS + \sum_{i=1}^{n} \frac{\partial E}{\partial V_i} dV_i.$$

These partial derivatives are given conventional names: $T := \partial E/\partial S$ is called the **temperature** at $x$, and $P_i := -\partial E/\partial V_i$ is called a (generalized) **pressure**. (The "honest" pressure is $-\partial E/\partial V$, where $V$ is the volume. The minus sign is conventional.)

These names are very suggestive, and it's worth explaining how they line up with the way you expect things with these names to behave. Suppose you have two systems at different temperatures $T_1$ and $T_2$, and you bring them into thermal contact with each other, so that they are able to exchange energy but all of the $V_i$'s stay fixed. Allow them to come to equilibrium. Write $E_1, E_2, S_1, S_2$ for the energies and entropies of the two systems.

The total energy $E = E_1 + E_2$ is conserved — we're assuming the two systems can't exchange energy except with each other — so we conclude that

$$\frac{\partial S}{\partial E_1} = \frac{\partial S_1}{\partial E_1} + \frac{\partial S_2}{\partial E_1} = \frac{\partial S_1}{\partial E_1} - \frac{\partial S_2}{\partial E_2} = \frac{1}{T_1} - \frac{1}{T_2}.$$

If $T_2 > T_1$, we see that moving energy from the second system to the first would increase the entropy, and vice versa if $T_1 > T_2$. In order for the entropy to be maximized, the tempetatures of the two systems must be equal.

So, using the strong version of the second law mentioned earlier, once the two systems have reached equilibrium with each other, *the temperatures have to be the same*. If, as we have agreed

to assume, the entropy is a concave function of the energy at fixed values of the $V_i$'s, then we have the stronger conclusion that in the process of equilibrating, *energy must flow from the system with higher temperature to the one with lower temperature*.

Concavity is useful for another reason. It's often convenient to be able to switch which variables you are using to parameterize the state space, and so it is helpful if, for example, each energy corresponds to exactly one temperature. This follows from the fact that entropy is a concave function of energy, because then $1/T = \partial S/\partial E$ is monotonic. Because $\partial^2 S/\partial E^2 = (-1/T^2)(\partial T/\partial E)$, we see that the concavity implies that temperature increases with energy, as one might expect. Similarly, it implies that decreasing the volume should increase the pressure.

Again, in general, these assumptions can be violated. In the presence of phase transitions, the argument we gave for concavity breaks down, there might not be a one-to-one correspondence between energies and temperatures, and there are systems one could write down for which temperature can be negative or can decrease with energy. I hope to talk about all of this, especially the theory of phase transitions, in a future article in this series, but for now, we will continue to assume that the entropy is a concave, increasing function of the energy.

The temperature of a system can be thought of as a measure of the tendency to spontaneously give off energy to anything it is in contact with. It's common to find less careful accounts of temperature that imply that it's somehow just "average energy in funny units," so it's important to emphasize that this is not even a little bit true. For instance, a kilogram of air has far less energy than a kilogram of water at the same temperature. (Part of the confusion, I think, stems from the fact that for an ideal gas there is a linear relationship between the kinetic energy per particle and the temperature. But this is just a fact about ideal gases, not a definition of temperature! It is not true at all for other types of systems, and even for different ideal gases the constant of proportionality can change.)

This picture also gives a good operational way to *measure* temperature: if we find some small system (like the mercury in a thermometer) for which some visible thermodynamic variable changes with temperature, we may bring it into thermal contact with the system we want to measure, wait for them to come to equilibrium, and read the value of the other, visible variable.

As for the identification of pressure with $-\partial E/\partial V$, imagine a gas in a box, one of the walls of which is a piston that can move in and out, changing the volume of the box. Suppose the surface area of the piston is $A$. Now, imagine that we push on the piston by applying a force $F$, slowly enough so that the entropy doesn't change (more on this assumption later), moving it inward by a distance $dx$. We have changed the volume by $-Adx$ and done *work* on the gas in the amount $Fdx$, and our assumptions imply that this is the only change in the energy, so

$$Fdx = \frac{\partial E}{\partial V} dV = -\frac{\partial E}{\partial V} A dx,$$

and we conclude that $-\partial E/\partial V$ is the force per unit area, the usual definition of pressure.

## 2.4 Heat and Work

Putting the definitions of temperature and generalized pressure back into the formula that led us to define them, we get

$$dE = T\,dS - \sum_{i=1}^{n} P_i\,dV_i.$$

Remember that this is a statement about the relationship between different *equilibrium states* of the system, since that's what is represented by points in the thermodynamic state space, *not*

a general formula for what happens when the state of a system changes in time. We can, though, draw a curve to represent the trajectory of a system over time if the change is **quasistatic**, which means that the change is slow enough that the system comes back to equilibrium over the course of the change faster than any appreciable change can occur. In this situation we can more or less safely model the system as though it's just in equilibrium the whole time, even though of course it must leave equilibrium a little bit in order for the state to change at all.

When moving quasistatically from one equilibrium state to another, it is common to refer to the first term as **heat** ($Q$) and the sum of the rest of the terms as **work** ($W$). (Note that the words "temperature" and "heat" mean different things in thermodynamics: heat is a form of energy; temperature is not!) It can sometimes be useful to separate out heat in this way, as the energy that has moved "due to" a difference in temperatures, and we just saw in our discussion of pressure that it can also be worthwhile to identify the second part with work. A lot is made in thermodynamics textbooks of the fact that, while $dE$ is an exact 1-form, the heat and work individually are not; in particular there's no quantity corresponding to the "amount of heat in a system."

This version of the formula — relating the change in energy to heat and work — is sometimes referred to as the first law of thermodynamics. I think it's usually less confusing to just not take heat and work seriously as "fundamental" notions and think first about energy and entropy, worrying about whether we want to divide up changes of energy like this only later. From this perspective, the first law really is just conservation of energy, and the other formula constitutes the definition of temperature and pressure.

The presentation I've chosen here centers energy and entropy and derives all the other interesting thermodynamic quantities from them, but this is not how these ideas arose historically. A more historical presentation would define temperature *operationally* using the procedure alluded to earlier where we equilibrate with a thermometer system. (The zeroth law would then allow us to argue that this is well-defined.) The division of the change of energy into heat and work is similarly practical: work is energy that goes toward moving macroscopic objects around — that is, changing the values of what we called the $V_i$'s — and heat is the portion of the change in energy that isn't attributable to work. (In particular, energy transferred between systems at different temperatures while holding all the other variables fixed is all heat, since by hypothesis no work is being done.)

There are versions of the second law that don't directly mention entropy. "Kelvin's second law" says that there is no cyclic process whose only effect is to convert some amount of heat into work; "Clausius's second law" says that there is no cyclic process whose only effect is to move heat from a cold reservoir to a hot one. One can then prove our version of the second law in two steps. First, one can show from either of these statements that, for a quasistatic change of state, $Q/T$ is an exact 1-form. We can use this to *define* entropy (up to an additive constant) by setting $dS = Q/T$. One then shows that if this $S$ could decrease, one could construct a cyclic process violating whichever version of the second law was chosen. This is the presentation followed in Thompson's book, which I recommend if you are interested in it.

Entropy defined in this way is only fixed up to an additive constant, and this constant can be fixed by making a choice for the value of $S$ at a single point. The **third law of thermodynamics** says that entropy of any system at a temperature of zero is the same, regardless of the details of the system. (Conventionally, this system-independent value is taken to be 0.) The third law is not especially relevant to anything else we'll do in this article — in fact, it doesn't seem to come up much at all — so we won't discuss it any further here.

I mentioned earlier that we can identify the heat with $T\,dS$ when the change is quasistatic. When non-equilibrium processes are involved, this identification can fail if we also want to

take "heat" to mean "$dE$ minus work." There is a nice example of this that I'm stealing from Scroeder's book. Imagine a gas in a box with a piston on one end. If you move the piston very quickly, faster than the typical speed of the gas particles, some of the particles will bunch up behind the piston and push back on it, requiring you to push harder to get the piston to move. If we move the piston a short distance in this violent way, and then allow the gas to come to equilibrium again, we find that the work we had to do was *more* than $-PdV$, and the heat is therefore less than $TdS$. This process has, in other words, created more entropy than can be accounted for just by whatever heat was transferred at the same time. Because it's impossible to get the volume to change without doing *at least* $PdV$ work, we always have $Q \leq TdS$.

## 2.5 Free Energy and Enthalpy

The second law of thermodynamics says that entropy cannot decrease in an isolated system, but most real systems are not anywhere close to being isolated. In a situation like this, where the entropy of the system alone won't let you usefully apply the second law, it helps to keep track of a slightly different set of thermodynamic variables. For simplicity, we're going to assume throughout this section that there is only one "additional" thermodynamic variable, the volume $V$.

Let's first consider a system which is **mechanically isolated**, that is, prevented from doing any work, but which is allowed to exchange energy with some environment that is so large that its temperature $T$ doesn't change appreciably when it exchanges energy with the system. (Such an environment is called a **heat bath**.) Since the whole point of this exercise is to keep track of the system in the process of equilibrating, we can't assume that it's in equilibrium over the course of this process, and so we can't apply the formula $dE = TdS - PdV$ from the previous section. (If you like, imagine that it's composed of several subsystems in the process of coming into equilibrium with each other as well as with the environment; these subsystems are free to transfer energy among themselves as long as the system as a whole remains mechanically isolated.)

In the language of the previous section, since no work is done, any energy transferred between the system and the environment takes the form of heat. We will assume that the environment equilibrates faster than it takes for an appreciable amount of heat to be transferred in this way (i.e., the state of the environment changes quasistatically) so if some infinitesimal amount of heat $dE_{\text{env}} = -dE_{\text{sys}}$ moves from the system to the environment, we have that $dS_{\text{env}} = dE_{\text{env}}/T$. (That is, because the environment, unlike the system, is in equilibrium the whole time, we *are* free to apply the formula for $dE$ to it.) But the *total* entropy of the system and the environment can't decrease, which means that

$$dS_{\text{sys}} \geq -dE_{\text{env}}/T = dE_{\text{sys}}/T.$$

So, if we define the **Helmholtz free energy** of the system as $F = E - TS$ the above inequality can be written

$$\frac{dF_{\text{sys}}}{T} \leq 0,$$

and we therefore conclude that for a mechanically isolated system in contact with a heat bath at constant temperature, the Helmholtz free energy cannot increase. This is our "replacement" for the second law in this case.

Let's now relax the constraint that $T$ is constant. We can write a useful expression for the

change in the equilibrium value of $F$ directly from the definition:

$$\begin{aligned} dF &= dE - T\,dS - S\,dT \\ &= -S\,dT - P\,dV. \end{aligned}$$

This is like the formula for $dE$, except that we have interchanged the $S$ and $T$ variables in the formula and introduced a minus sign on that term. Compare the expressions for thermodynamic variables in terms of partial derivatives that we get from the formulas for $dE$ and $dF$:

$$T = \left.\frac{\partial E}{\partial S}\right|_V \qquad S = -\left.\frac{\partial F}{\partial T}\right|_V$$

$$P = -\left.\frac{\partial E}{\partial V}\right|_S \qquad P = -\left.\frac{\partial F}{\partial V}\right|_T.$$

(Here we have written which variables are held constant in a partial derivative using a subscript on the right.) The procedure that turns $E$ into $F$ is a simple example of a **Legendre transform**; this is also how we move between velocity and momentum coordinates when moving between Lagrangian and Hamiltonian mechanics.

The name "free energy" comes from analyzing the situation when the system is not mechanically isolated, but is still in contact with a heat bath at constant temperature $T$. Since any energy that is transferred in the form of heat is (by definition) useless for moving macroscopic objects around, we'd like to know how much *work* can be done during this process. We have $dF = dE - T\,dS = Q + W - T\,dS$. I encourage you to repeat the analysis that began this section and conclude that $Q \leq T\,dS$, which implies that $dF \leq W$.

In particular, when both $dF$ and $W$ are negative, we conclude that the amount of work that can be performed by the system (while keeping the temperature constant) is bounded by the change in its Helmholtz free energy, so the $dF$ energy is "free" during this process in the sense of being available to do work. Conversely, just absorbing heat can increase $E$, but I need to do *work* on the system to increase $F$.

A similar analysis carried out for a system in an environment at constant temperature and pressure (but whose volume can change in addition to its energy) leads us to define

$$G = E - TS + PV,$$

the **Gibbs free energy**, which is the Legendre transform of $E$ with respect to both the entropy/temperature and volume/pressure pairs of variables. A similar conclusion about the second law applies, as does a similar conclusion about the amount of work that can be done if we also count the $P\,dV$ contribution to $dE$ as "useless" along with $Q$; in a constant-pressure environment, the expansion and contraction of the container happens "automatically" in just the same way as heat transfer in a constant-temperature environment.

The Legendre transform of $E$ with respect to just the volume/pressure variables is written

$$H = E + PV,$$

and it's called **enthalpy**. The enthalpy is a useful variable to keep track of in settings (chemical reactions are a common example) where you are interested in keeping track of the movement of energy and you have control of the pressure but not the volume of the system; in such a situation the system might do work on its surroundings in the process of expanding, and it helps to keep track of something that's insensitive to such a change.

## 2.6 The Carnot Cycle

We'll close this section by considering a classic application of the ideas we've built so far: an analysis of how much energy you can extract from a **heat engine**. A heat engine is any device that extracts energy from two large systems at different temperatures and uses it to perform work. We will be looking at a particular process for accomplishing this called the **Carnot cycle**, but many of our conclusions will apply equally to all heat engines.

The setup for a heat engine consists of three pieces: some large amount of some hot substance at temperature $T_{\text{hot}}$, called the **hot reservoir**; a **cold reservoir** at temperature $T_{\text{cold}}$; and a smaller amount of gas called the **working fluid**. By repeatedly exposing the working fluid to the two reservoirs to change its temperature, we will use the resulting changes in its volume to perform work. (You can imagine that one of the walls of the working fluid's container is a piston that the fluid pushes on when it expands, and the piston is attached to some object you want to move.) To keep the expected relationships between the thermodynamic variables straight, it can help to imagine that the working fluid is an ideal gas, so that $PV = NkT$ always, but this isn't required for the analysis to work.

Write $Q_{\text{hot}}$ for the heat transferred from the hot reservoir to the working fluid during a cycle, $Q_{\text{cold}}$ for the heat transferred from the working fluid to the cold reservoir, and $W$ for the total net work performed by the engine over a whole cycle. We define the **efficiency** of the engine as the work extracted divided by the energy we need from the hot reservoir, that is,

$$e = \frac{W}{Q_{\text{hot}}} = \frac{Q_{\text{hot}} - Q_{\text{cold}}}{Q_{\text{hot}}}.$$

We assume that the two reservoirs have fixed volumes, which means that the entropy of the hot reservoir changes by

$$\Delta S_{\text{hot}} = -\frac{Q_{\text{hot}}}{T_{\text{hot}}},$$

and similarly with the opposite sign for the cold reservoir, since the only change in the reservoir's energy comes from the $TdS$ term and $T$ is presumed to be constant. The working fluid must return to its original state at the end of a cycle (this is a hypothesis of this whole setup) so in particular its entropy doesn't change.

The process of going through a cycle might create some entropy in the environment surrounding the engine, though. We therefore from this computation and the second law that

$$\frac{Q_{\text{cold}}}{T_{\text{cold}}} - \frac{Q_{\text{hot}}}{T_{\text{hot}}} \geq 0,$$

which after a quick computation yields a bound on the efficiency:

$$e \leq 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}},$$

with equality if and only if the total entropy stays the same over the course of a whole cycle.

The Carnot cycle serves as an existence proof that getting arbitrarily close to this bound is possible, at least in principle. (In practice, a Carnot engine would run so slowly as to be basically useless, but it would be *efficient* the sense just defined.) Every step is assumed to be quasistatic, allowing us to use the formula $dE = TdS - PdV$ from earlier.

The Carnot cycle consists of four steps, repeated, as the name suggests, in a cycle:

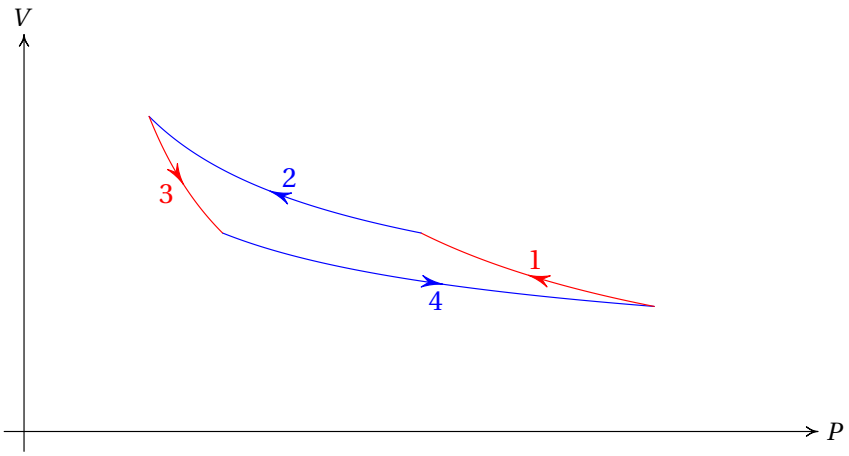
1. With the temperature of the working fluid just slightly below $T_{\text{hot}}$, place it in contact with the hot reservoir and allow energy to flow from the hot reservoir to the fluid. In order to keep the temperature of the gas from changing, we allow its volume to expand. This process is called **isothermal expansion**. The working fluid has absorbed some heat from the hot reservoir and converted some but not all of that energy to work.

2. Next, we disconnect the working fluid from the hot reservoir, and we allow it to expand some more. Since the working fluid is no longer absorbing any heat, this has the effect of lowering its temperature. We assume that this is done at constant entropy, a reasonable assumption if the fluid is thermally isolated during this process and the movement of the piston is frictionless. This step is called **isentropic expansion**. We do this until the working fluid's temperature is just above $T_{\text{cold}}$.

3. Next, we place the working fluid in contact with the cold reservoir and allow it to contract, so that its temperature stays the same. This is **isothermal compression**.

4. Finally, in order to get the gas back to its original state, we disconnect the working fluid from the cold reservoir and allow it to compress some more, until its temperature and volume are both back to their original values. This is **isentropic compression**.

It's common to draw a picture of how $P$ and $V$ change over the course of these steps on a so-called "$PV$-diagram." This is what the four steps look like for an ideal gas.




The red and blue lines are lines of constant $T$ and constant $S$, called isothermal curves and isentropic curves respectively. (Note that, while the entropy of the system and the reservoirs *together* never changes over the course of a cycle — another way of saying that the Carnot cycle is *reversible* — the entropy of the *system alone* does change during the isothermal expansion and compression phases as it exchanges heat with the reservoirs.) To draw the isentropic curves, I made use of the formula $E = \frac{3}{2}NkT$ for the energy of a monatomic ideal gas, which we will also derive in the statistical mechanics section; it's a nice exercise to see how this produces a formula relating $P$ and $V$ in the isentropic case.

During the expansion phases the working fluid does work, and during the compression phases work is done on the working fluid. But, since the pressure is higher during the expansions than the compressions, the net work done by the fluid is positive. This necessarily means that the fluid absorbs more heat from the hot reservoir than it gives to the cold reservoir. The exact

amount of work is the integral of $PdV$ along the curve in the diagram, which by Green's theorem is equal to the area of the region it encloses.

If, in fact, the temperature of the working fluid is close enough to the temperatures of the reservoirs that no appreciable entropy is created there, and if the other expansion and compression steps are in fact isentropic, then entropy will in fact be completely conserved over the course of a whole cycle. In this (idealized) situation, the inequality we calculated above will actually be an equality, and so the efficiency of the engine will in fact meet the bound.

At any rate, though, if we run the Carnot cycle over and over, the net effect is that heat flows from the hot reservoir to the cold reservoir, and unless more energy is being poured into the reservoirs from the outside this will shrink the difference between their temperatures. (We've modeled the reservoirs as being so large that their temperatures don't change when they give off or absorb heat, but this is only an approximation.) The engine is only useful as long as this temperature difference can be maintained.

# 3 Statistical Mechanics

We now switch our focus to a discussion of *statistical mechanics*. This is the framework that describes how thermodynamics arises from the microscopic laws of physics, but it would be a mistake to think of that as its only function, or even its primary function. A description of a system on the level of statistical mechanics is much more informative than a thermodynamic description, and the success of the statistical-mechanical framework rests on the fact that the details of this description — beyond the mere fact of thermodynamic behavior — are themselves well-confirmed by experiment.

As a very simple example, we will see that the statistical-mechanical machinery allows us to prove the ideal gas law as a theorem. But this should be seen as just the beginning. While here we'll mostly be concerned with just setting up the machinery, I hope to explore its consequences much more in a future article in this series.

## 3.1 States as Distributions on Phase Space

Our presentation of statistical mechanics will be built out of classical, nonrelativistic Hamiltonian mechanics. (There is also a theory of quantum statistical mechanics, which we won't touch on in this article.) We'll very briefly recall how this theory works; there is an article in this series you can read for more details.

The state of a system corresponds to a point in *phase space*, which we will denote by $X$ and which is usually the cotangent bundle of *configuration space* $Q$. We will assume that $Q$ is a compact manifold; imagine, for example, a gas confined to a finite box. Phase space has the structure of a symplectic manifold, and in the cotangent bundle case, if we have local coordinates $q_i$ and corresponding coordinates $p_i$ on the cotangent spaces, the symplectic form is $\omega = \sum_i dp_i \wedge dq_i$. The dynamics — the rules for how the system evolves in time — are determined by a real-valued function on $X$ called the *Hamiltonian* $H$. Using the symplectic form, we may turn $dH$ into a vector field and the resulting *Hamiltonian flow* produces the dynamics. This flow preserves both $\omega$ and $H$. As a function of the state, $H$ is the total energy, and so we see that energy is conserved.

The phase space for a macroscopic system has a truly enormous number of degrees of freedom; it's not even remotely practical to learn where every single particle is at any time. Instead, in statistical mechanics we represent our knowledge of the state of the system as

a *probability distribution* on $X$. As in our discussion of thermodynamics, we will mostly be concerned with *equilibrium statistical mechanics*, and so our main task will be to find probability distributions we can use to represent an equilibrium state and extract the thermodynamic variables like entropy, temperature, and so on from it.

On any symplectic manifold of dimension $2n$ we can produce a volume form, and therefore a measure, by taking $\omega^{\wedge n}$. The resulting measure is called the **Liouville measure**, which we'll write $\mu^L$. If $p_1, \ldots, p_n, q_1, \ldots, q_n$ are local coordinates in which $\omega = \sum_i dp_i \wedge dq_i$, then we may use the $p_i$'s and $q_i$'s to pull back the Lebesgue measure from $\mathbb{R}^{2n}$, and it will coincide with the Liouville measure. Because Hamiltonian flows preserve $\omega$, they also preserve $\mu^L$, that is, the Liouville measure is preserved by time translation. This result is often called *Liouville's Theorem*. The Liouville measure will play a crucial role in our construction of equilibrium distributions.

The use of probability distributions also neatly addresses another possible difficulty. The task of extracting thermodynamics from the microscopic laws of physics seems to face an insurmountable problem: the laws of physics have a time-reversal symmetry, but in thermodynamics the approach to equilibrium and increasing entropy happen in only one time direction. This objection forces us to slightly weaken our claim. We can't claim that it's *impossible* to end up in a state with lower entropy — after all, you can get a path with this property simply by reversing one in which entropy increases — instead, we claim it's very *improbable*. The time symmetry is then broken by the fact that our initial measurements of the system constrain our knowledge of the initial state of the system rather than the final state. (There is much more on the topic of how this resolves the problem in the companion piece.)

Not every imaginable system behaves thermodynamically, and so this equilibrium-seeking behavior can't somehow follow directly from Hamiltonian mechanics on its own. Ideally, we would be able to list some reasonable conditions on the Hamiltonian and use them to give a completely rigorous account that goes directly from Hamiltonian mechanics to a proof of thermodynamic behavior with no gaps. Unfortunately, even in cases where we expect it to happen, this seems wildly out of reach. While it's possible to provide such an account in some very simple, unrealistic models, most of the time we *assume* that the dynamics are such that equilibrium exists and that (under some suitable probability distribution) the vast majority of initial states end up there.

The story about how this is supposed to happen has two parts. The first is the claim that, after we have fixed our list of thermodynamic variables and our probability distribution, an overwhelmingly large fraction of the states will have values for the thermodynamic variables lying in a very small range. We therefore refer to the expected value of each variable as its "equilibrium value."

This half of the story can be established rigorously in some cases. A helpful picture is to imagine dividing the system into a large number of small pieces (but still much larger than a single particle). It is often the case that the thermodynamic variable in question can be written as a sum over all the pieces of some quantity that depends only on each piece, plus a small error term. If you can do this, and if these per-piece quantities are sufficiently close to independent under your chosen equlibrium distribution, then the law of large numbers should lead you to expect this sharply peaked behavior. There are many results which, under certain assumptions on the form of the Hamiltonian, prove rigorous bounds of this form, but we won't go over them here. The standard reference for this is the book *Statistical Mechanics: Rigorous Results* by David Ruelle. (I found it clearly written, but be aware that it was published in 1969.)

The second part of the story is that, if the dynamics are "chaotic" enough that most states get jostled around phase space more or less randomly, an arbitrary state is likely to eventually end up in the large region of phase space where the values of the variables are close to their

equilibrium values, and a state in this large region is likely to stay there. In particular, the time symmetry we discussed earlier is broken not by the laws of motion but by the initial condition: if we assume that the system starts in a state with *a priori* unlikely values for the thermodynamic variables, we conclude that they will move toward their equilibrium values simply because that is where the vast majority of states end up no matter what. Unfortunately, proving that this sort of behavior actually occurs seems completely out of reach in any realistic model, and so we are forced to just assume it.

Throughout this section, we will often refer to both probability measures (denoted by some form of the symbol $\mu$) and probability densities (some form of $p$). Unless indicated otherwise, the densities will always be densities with respect to the Liouville measure, that is, when we say that some measure $\mu_i$ has density $p_i$, we mean

$$\mu_i(A) = \int_A p_i(x) d\mu^L(x).$$

## 3.2 The Microcanonical Distribution

Our main task in setting up equilibrium statistical mechanics is to choose probability distributions to represent a system at equilibrium. The first equilibrium distribution we will consider will be for an *isolated* system, that is, one which is completely cut off from its environment, and so in particular can't exchange any particles or energy with anything else. For such a system, the total energy is exactly conserved, so whatever distribution we end up using will be supported on some constant-energy hypersurface $\Sigma_E := \{x \in X : H(x) = E\}$. We will assume throughout that $\Sigma_E$ is compact, an assumption made more reasonable by the assumed compactness of $Q$.

An equilibrium distribution ought to be time-symmetric, that is, to be preserved when time is run forward. The Liouville measure has precisely this property, so we can use it to build a measure on $\Sigma_E$: restrict the Liouville measure to $\{x \in X : E - \Delta E \leq H(x) \leq E + \Delta E\}$, divide by $2\Delta E$, and let $\Delta E$ go to zero. We can call the resulting measure on $\Sigma_E$ the **restricted Liouville measure**. Because $\Sigma_E$ is compact, its total measure will be finite, and so we can build a probability distribution by simply dividing by this total measure. We call the result the **microcanonical distribution of energy** $E$, which we'll write $\mu_E^m$.

If we work in coordinates in which the symplectic form looks like $\sum_i dp_i \wedge dq_i$, this measure does *not* just give the "surface area" of a subset of $\Sigma_E$. Rather, I encourage you to show that the measure of some subset $A \subseteq \Sigma_E$ is given by

$$\mu_E^m(A) = \frac{1}{\Omega_E} \int_A \frac{d^{n-1}x}{|\nabla H|},$$

where $\Omega_E = \int_{\Sigma_E} d^{n-1}x/|\nabla H|$ is the total measure of $\Sigma_E$ under the restricted Liouville measure, and the surface area measure $d^{n-1}x$ and the gradient $\nabla H$ are computed in the given coordinates.

A natural question to ask at this point is to what extent this particular distribution is "forced" on us. Are there other distrbutions that are preserved by the dynamics that we could have used instead? Because energy is conserved, we can multiply the Liouville measure by a function of $H$ and the resulting measure would also be preserved by the dynamics. But I encourage you to check that this actually would not change the microcanonical distribution at all.

There is one more interesting case, though: there might be some conserved quantity other than the energy that we have failed to keep track of. (For example, for a gas in a perfectly cylindrical container, we would need to think about the angular momentum about the central

axis.) In this case, the surfaces on which that quantity takes a constant value will be separately preserved, and so we would be free to multiply $\mu_E^m$ by any function which only depends on the value of the conserved quantity. In such a situation, we might also fix the value of that quantity as well, and build in an analogous way a distribution supported on the surface on which both the energy and this new quantity are fixed.

For simplicity, we'll assume for the rest of this discussion that there are no such extra conserved quantities to worry about. Even in this case, there is no general proof that $\mu_E^m$ is the unique probability measure on $\Sigma_E$ preserved by the dynamics. In fact, as far as I know, there is no airtight argument that the microcanonical distribution is the only "correct" one to use to describe our situation, nor even complete agreement about what such an argument would even entail. I think it's best to regard the choice of the microcanonical distribution as a *postulate* of statistical mechanics. It's one of the building blocks of the theory, and we can test the theory against experiment to see how well it describes reality.

Earlier, when describing why it is plausible that most systems will approach equilibrium, we said that we will assume that, for most states, the values of thermodynamic variables will be very close to their equilibrium values. This amounts to assuming that, under the microcanonical distribution, those variables are very sharply peaked around their expected values. (In other words, we are using the microcanonical distribution to decide what "most" means.) Our assumption about the approach to equilibrium then amounts to the claim that, if we start with some other distribution and evolve it forward for a long enough time, then the expected values of our variables will tend toward their expected values under the microcanonical distribution, and their variances will become small.

## 3.3 Entropy and Information Theory

It's relatively straightforward to see how energy is supposed to emerge from this statistical-mechanical framework: for an individual point in phase space, it's the same concept as in Hamiltonian mechanics, and to a probability distribution we can assign the expected value of this same quantity. Many thermodynamic variables, like angular momentum, volume, or the number of particles, can be identified with an expected value in the same way.

But entropy is different: in our framework, entropy will *not* be a property of an individual point in phase space but of a probability distribution as a whole. (This is therefore also true of quantities that are derived from entropy, like temperature.) The quantity we'll use to represent thermodynamic entropy is, in fact, almost identical to the quantity called "entropy" in information theory, so we'll give a lightning-fast review of this concept now. This may not be enough if the concept is brand new to you. I encourage you to seek out a more detailed explanation elsewhere in such a case.

We'll first consider probability distributions on a finite set. Let $\Omega$ be a finite set and consider a probability distribution $p$ on $\Omega$, which amounts to a nonnegative real number $p_i$ for each $i \in \Omega$ with $\sum_i p_i = 1$. For each $i$, we say that the **surprisal** of the result $i$ under $p$ is $-\log p_i$. It's useful to think of this as representing how much information you have gained when you take a random sample from $p$ and see that it is $i$. The logarithm is there to make it additive for independent samples.

The **entropy** of $p$ is then the expected value of the surprisal:

$$S_{\text{info}}[p] = -\sum_{i \in \Omega} p_i \log p_i,$$

where the convention is that if some $p_i = 0$ then it contributes zero to the sum. The entropy

should be thought of as measuring of how much information, on average, you gain when you learn the identity of a random sample from $p$. This means it's a measure of *ignorance*: lower entropy means knowing that the data are distributed according to $p$ is already very informative, so there is not much more you can learn when you see a new sample. The lowest-entropy distribution on $\Omega$ is the one concentrated at a single point, which has entropy 0; highest is the uniform distribution, which has entropy $\log |\Omega|$.

There is an extra complication that arises for continuous probability distributions. If $p$ now represents a probability density, it's tempting to define the entropy as $-\int_\Omega p(x) \log p(x)$. But sadly this can't work: a probability density is only well-defined with respect to a background measure, and the choice of measure will affect the value of this integral. (A good sanity check is that while probabilities are unitless, probability densities have units of inverse volume, and so it is inappropriate to take their logarithms.) Without making any additional choices, there is no coherent way to extend the concept of entropy to the continuous setting.

If we permit ourselves to fix a measure $\mu^B$ in the background, though, we can define the **relative entropy** of a measure $\mu$ with respect to $\mu^B$ as

$$S_{\text{info}}[\mu||\mu^B] := -\int_\Omega \log\left(\frac{d\mu(x)}{d\mu^B(x)}\right) d\mu(x).$$

(You may have seen this definition without the minus sign, especially under the name "Kullback-Leibler divergence." It is conventional to include it here so that, when we eventually discuss the second law, larger entropies still have the same meaning as in thermodynamics.) Here $d\mu/d\mu^B$ denotes the Radon-Nikodym derivative; if we are given a third "reference" measure $\mu^R$ with respect to which both $\mu$ and $\mu^B$ are absolutely continuous, we can also write this as

$$-\int_\Omega p(x) \log\left(\frac{p(x)}{p^B(x)}\right) d\mu^R(x),$$

where $p$ and $p^B$ are densities with respect to $\mu^R$.

It's again useful to think of this in information-theoretic terms: if the background measure is taken to represent the position of total ignorance, then the relative entropy represents how much information we gain on average when we see a sample from $\mu$. For the same reason as in the finite case, it is useful to think of the low-entropy distributions as the more "informative" ones. Unlike entropies on finite sets, the relative entropy is always *nonpositive*, and reaches its maximum value of 0 exactly when $\mu = \mu^B$.

It's common to use Bayesian language to talk about this situation, referring to $\mu^B$ as a "prior." This is fine as long as you allow your class of priors to include measures for which $\mu^B(\Omega) = \infty$, so-called **improper priors**. While such a $\mu^B$ can't really be thought of as a belief about how likely some subset of $\Omega$ is to arise, you might think of it as specifying *ratios* of such likelihoods. (In particular, then, an improper prior implies a belief about a relative probability if we are conditioning on a set of finite measure.) For example, using the Lebesgue measure on $\mathbb{R}$ as an improper prior means expressing the belief that, in the absence of other information, the likelihood for a sample to land in some interval should be proportional to the interval's length.

Note that if $\mu^B = \mu^R$, then $p^B = 1$ and the second integral above will look exactly like the one we just said was invalid! So it's fine to write that expression so long as you remember that it is actually a relative entropy in disguise, and in a setting where the choice of background measure is understood it's common to be a bit sloppy with language and just call it the "entropy." We will, in fact, work in such a setting: *whenever we refer to entropies from now on, we are always actually talking about relative entropies with respect to the Liouville measure.*

## 3.4 Marginalizing the Microcanonical Distribution

The microcanonical distribution is simple to write down, but it has a couple of disadvantages that make it difficult to use in actual computations. First, the fact that it's supported only on the hypersurface $\Sigma_E$ turns out to make it hard to work with computationally. Second, perhaps most importantly, the assumption we started with — that the system never exchanges energy with its environment — is physically unrealistic. We would therefore like a distribution that is suitable for describing a *non-isolated* system at equilibrium with its environment.

We can learn a lot about what properties we'd like our distribution to have by examining this second problem in more detail. We'll model the physical situation by splitting the phase space into two pieces, which we'll call the *system* and the *environment*, so that $X = X_{\text{sys}} \times X_{\text{env}}$. We'll assume that the environment is much larger than the system, and that, while they can interact, the energy of this interaction is much smaller than the energy of either part separately. (This is a good assumption for the type of situation we're usually interested in modelling, where the interaction occurs along some interface whose size grows like an area while the size of the system itself grows like a volume.) This means that we can express the Hamiltonian in the form

$$H = H_{\text{sys}} + H_{\text{env}} + H_{\text{int}},$$

where $H_{\text{sys}}$ depends only on $X_{\text{sys}}$, $H_{\text{env}}$ depends only on $X_{\text{env}}$, and $H_{\text{int}} \ll H_{\text{sys}} \ll H_{\text{env}}$.

Now, suppose the state of the system and the environment taken together is distributed according to the microcanonical distribution, but we are interested in the state of the system alone. We want to *marginalize* the microcanonical distribution, that is, to take the measure on $X_{\text{sys}}$ defined by $\mu(A) = \mu_E^m(A \times X_{\text{env}})$. Because the interaction term in the energy is so small, we work in the approximation in which the total energy is simply the sum of the energies of the system and the environment. Even if we do this, though, the marginal distribution is not itself a microcanonical distribution, because the state of the system is not restricted to any constant-energy hypersurface. (In fact, assuming all our Hamiltonians are bounded below by 0, the energy of the system can be anything between 0 and $E$.)

Fortunately, in reasonable cases there is a family of distributions which (a) is closed under marginalizations of this type, where the total energy is just the sum of the energies of the system and the environment, (b) agrees with the microcanonical distribution in the limit as the number of particles goes to infinity, and (c) can be described explicitly. We define the **canonical distribution of average energy** $E$ to be the distribution which, among all distributions in which the expected value of the energy is $E$, maximizes the entropy

$$S_{\text{info}}[p] := -\int_X p(x) \log p(x) d\mu^L(x),$$

where $p(x)$ is the density with respect to the Liouville measure.

(This condition might or might not pick out a unique probability distribution, or even pick out any distribution at all; in many cases it does, and we will proceed for now under this assumption, but the language "the canonical distribution" is reserved for the cases in which it is true. It will also turn out that, unlike for the microcanonical distribution, the average energy is not actually the most natural parameter to use for the canonical distribution. We'll discuss both issues more momentarily.)

The defining condition of the canonical distribution is a *maximum entropy condition*. Again, we are actually maximizing *relative* entropy with respect to the Liouville measure, and this gives us a useful way to interpret the condition: if the Liouville measure represents complete ignorance

about the state, then the canonical distribution is the one which, among the distributions with average energy $E$, is *maximally uninformative*, that is, which assumes as little additional information as possible.

But whatever interpretation you want to attach to it, the maximum entropy condition can be used to establish conditions (a) and (b) above.

The proof of the first condition — that canonical distributions are preserved by marginalization — is the more straightforward of the two. We'll make use of the following fact. Suppose $\mu$ is a probability distribution on $X_1 \times X_2$; write $\mu_1$ for the marginal distribution on $X_1$ and similarly for $\mu_2$. Then $S_{\text{info}}[\mu] \leq S_{\text{info}}[\mu_1] + S_{\text{info}}[\mu_2]$, with equality if and only if $\mu$ is the product distribution of $\mu_1$ and $\mu_2$. So, now suppose $\mu$ is the canonical distribution with average energy $E$, and assume as we have been that the energy is additive across $X_1$ and $X_2$. Then $\mu$ must be the product of $\mu_1$ and $\mu_2$, since otherwise we could replace it with the product and increase its entropy without affecting the expected value of the energy. But then $\mu_1$ and $\mu_2$ must be the maximum-entropy distributions for their respective average energies since, again, otherwise we could increase the entropy of $\mu$. We conclude that $\mu_1$ and $\mu_2$ are also canonical distributions.

The second fact — that the canonical and microcanonical distributions coincide in the many-particle limit — belongs to a collection of theorems called **equivalence of ensembles**; the word "ensemble" (which I have deliberately avoided using) is often used in statistical mechanics to refer to the collection of possible microscopic states from which we are sampling when working with one or the other of these distributions. I will refer you to Ruelle's *Statistical Mechanics: Rigorous Results* for proofs and just give a heuristic argument here.

Talking about what happens "as the number of particles goes to infinity" requires considering a *family* of systems with increasing values of $N$. For example, for a gas in the microcanonical distribution, we might fix a value $\rho$ for the density and $e$ for the average energy per particle and take the $i$'th system to be a gas with $N_i$ particles confined to some box of volume $N_i/\rho$ and total energy $N_i e$, and similarly for canonical distribution. We refer to the process of allowing $N$ to go to infinity in this way as taking the **thermodynamic limit**. The goal is then to show that, in the thermodynamic limit, some measure of the difference between the two distributions, like the relative entropy or the total variation distance, goes to zero.

(Ruelle's book does not, I think, prove exactly this statement about measures; the second paper by Touchette mentioned in the introduction shows how to extract it from the results that Ruelle does prove.)

Rigorous equivalence-of-ensembles results assume a specific form for the Hamiltonian, and in particular that the interaction between particles is *short-range* in a certain precise sense. This has the effect of making the energies of distant pairs of particles close to independent from each other under the canonical distribution. This enables us to make a law-of-large-numbers-like argument that, for a large number of particles, the energy is tightly peaked around its expected value. (This is also the source of the assumption we made back in the thermodynamics section about entropy being additive in composite systems: if the interaction is weak enough, then knowing the energy of one component tells you very little about the energy of the other, so they are close enough to independent to be treated as such, and as mentioned above, entropy is additive in the independent case.)

For any $S \subseteq X$ with $0 < \mu^L(S) < \infty$, the unique distribution of maximal entropy among those supported on $S$ is $1/\mu^L(S)$ times the restriction of $\mu^L$ to $S$. The microcanonical distribution arises by first taking such a maximum-entropy distribution supported on $\{x \in X : E - \Delta E \leq H(x) \leq E + \Delta E\}$ and then taking the $\Delta E \to 0$ limit. We have two conditions we might place on a distribution — that the energy be exactly $E$, or that the expected value of the energy be $E$ — but for a large number of particles, the second condition comes close to implying the first, so it is at

least plausible that the distributions would coincide in the thermodynamic limit.

This, then, is why we use the canonical distribution to describe the state of a non-isolated system. If we start with a microcanonical distribution for the combined state of the system and the environment, then as the size of the environment goes to infinity we are free to replace it with a canonical distribution. But once we've made this replacement, we see that the marginal distribution for the system alone is another canonical distribution, which is exactly what we needed.

## 3.5 A Formula for the Canonical Distribution

Our next task is to find a formula for the canonical distribution. This will be what enables us to finally connect statistical mechanics to thermodynamics, and in particular to see which quantities correspond to temperature, entropy, and the rest.

We're looking for a function $p$ which, among all functions for which

$$\int_X p(x)d\mu^L(x) = 1 \qquad \text{and} \qquad \int_X p(x)H(x)d\mu^L(x) = E,$$

maximizes the quantity

$$-\int_X p(x)\log p(x)d\mu^L(x).$$

We can solve this using an infinite-dimensional version of the Lagrange multiplier formalism. The formal statement we need is given quite concisely on this Wikipedia page; in our setting, it amounts to the fact that, given any function $p$ solving this constrained optimization problem, there exist constants $\beta$ and $\gamma$ so that $p$ also solves the *unconstrained* optimization problem of maximizing

$$\int_X [-p(x)\log p(x) - \gamma p(x) - \beta p(x)H(x)]\, d\mu^L(x).$$

We can now employ the standard calculus of variations trick: consider any smooth one-parameter family of functions $p_t(x)$ for which $p_0 = p$. If $p$ maximizes our functional, it must also be the case that

$$\begin{aligned} 0 &= \left.\frac{\partial}{\partial t}\right|_{t=0} \int_X [-p_t(x)\log p_t(x) - \gamma p_t(x) - \beta p_t(x)H(x)]\, d\mu^L(x) \\ &= \int_X \left.\frac{\partial p_t(x)}{\partial t}\right|_{t=0} (-\log p(x) - 1 - \gamma - \beta H(x))d\mu^L(x). \end{aligned}$$

In order for this to hold for *all* variations $p_t$, it must be the case that the quantity inside parentheses in the integral vanishes. (This is for the standard variational calculus reason: if the quantity inside parentheses is nonzero at some $x$, we can choose a variation which is only nonzero in a tiny neighborhood of $x$ and see that the corresponding integral will not vanish.) We conclude that $p(x) = \exp(-1-\gamma)\exp(-\beta H(x))$. It is standard to eliminate $\gamma$ by writing the distribution in the form

$$p^c_\beta(x) = \frac{1}{Z(\beta)}\exp(-\beta H(x)),$$

where

$$Z(\beta) = \int_X \exp(-\beta H(x))d\mu^L(x);$$

if we had left in the factors containing $\gamma$ they would cancel in this expression.

Given any value of $\beta$, we can use this formula for $p^c_\beta$ to compute the average energy $E$. But there is no guarantee that this process is invertible; there are Hamiltonians for which the map from $\beta$ to $E$ is neither injective nor surjective. This notably happens in the presence of *phase transitions*, which is a topic I hope to cover in a future article in this series. (This issue comes up in the "equivalence of ensembles" results we discussed above: part of showing that a given microcanonical distribution agrees with some canonical distribution in the thermodynamic limit is showing that we can associate a unique $\beta$ with each $E$ in exactly this way.) For now, though, we will assume that this problem does not occur; this is the case for many of the simplest systems one might analyze using this machinery, including the ideal gas that we'll discuss momentarily.

## 3.6 Thermodynamics from Statistical Mechanics

Given that we used the same symbol and the same name, it is probably no surprise that the information-theoretic entropy we have been discussing will end up serving the role of the entropy from thermodynamics. Conventionally, the two quantities are taken to differ by an affine transformation

$$S = kS_{\text{info}} + \text{constant},$$

where $k$ is Boltzmann's constant. Since observable quantities in thermodynamics only involve derivatives of $S$, the additive constant has to be fixed by other considerations, and we'll take this up in the next section. *For now, since it doesn't affect anything in this section, we will set this constant to zero.* We'll also see momentarily that a version of the second law applies for this $S$, but if we take this identification for granted for just a moment, we can see how the quantities we discussed in the thermodynamics appear.

(You might have seen a different definition of entropy, where we divide the phase space into regions of "macroscopically indistinguishable" states and define the entropy of a state to be $S_B = k \log W$, where $W$ is the volume of the region the state occupies. This is called the "Boltzmann entropy" and what we are using is called the "Gibbs entropy"; the Boltzmann entropy is, up to a constant, the Gibbs entropy of the uniform distribution on the region in question. The Boltzmann entropy has the advantage of being definable for an individual point in phase space once the regions have been chosen, but this rarely matters much; the Gibbs entropy is what is used to do most actual computations, so it's what we'll use too.)

The $Z$ appearing in the canonical distribution is called the **partition function**, and it contains a lot of information about the system. For example, I encourage you to check that

$$E = \int p(x)H(x)d\mu^L(x) = -\frac{d}{d\beta}(\log Z)$$

and

$$S = -k \int p(x) \log p(x) d\mu^L(x) = k(\log Z + \beta E).$$

These equations imply that

$$\frac{dS}{dE} = k\left(\frac{d}{dE}(\log Z) + \frac{d\beta}{dE}E + \beta\right) = k\left(\frac{d\beta}{dE}(-E) + \frac{d\beta}{dE}E + \beta\right) = k\beta,$$

which means that $\beta = 1/(kT)$, and so it is called the **inverse temperature**. A similar computation shows that $-(\log Z)/\beta$ is the Helmholtz free energy. Note that these formulas give us a

Legendre transform relation that is slightly different that the one we saw when we first discussed the free energy: $S$ is the Legendre transform of $\log Z$ with respect to the pair of variables $\beta, E$.

We derived the formula for the canonical distribution by imagining that our system is able to slowly exchange energy with its environment and concluding that we want the distribution which maximizes entropy for a fixed expected value of energy. We might want to also treat some quantity other than energy in this way at the same time.

In general, then, we can build an equilibrium distribution by specifying three types of thermodynamic variables:

- Variables *specified exactly*, like the energy in the microcanonical distribution.
- Variables *with a specified expected value*, like the energy in the canonical distribution.
- *Parameters* that the other variables (especially the Hamiltonian) might depend on. These might include the volume of the container, or something like the strength of an external magnetic field.

We then take the maximum-entropy distribution satisfying these constraints. In order for this to be an equilibrium distribution, the variables in the first two groups should be preserved by the dynamics. Suppose the variables in the second group are $A_1, \dots, A_m$ and $A_i$ is constrained to have expected value $a_i$. Using the Lagrange multipliers as above, we get:

$$p(x) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{m} \lambda_i A_i(x)\right)$$

$$Z(\lambda_1, \dots, \lambda_m) = \int \exp\left(-\sum_{i=1}^{m} \lambda_i A_i(x)\right) d\mu^L(x)$$

$$a_i = -\frac{\partial}{\partial \lambda_i}(\log Z)$$

$$S = k\left(\log Z + \sum_{i=1}^{m} \lambda_i A_i\right); \qquad \lambda_i = \frac{1}{k}\frac{\partial S}{\partial A_i}.$$

If energy is among the $A_i$'s, say $E = A_m$ and $\beta = \lambda_m$, then $-(\log Z)/\beta$ is the analogue of the Gibbs free energy, in which all the variables have undergone a Legendre transform. We say that the $A_i$ and $\lambda_i$ variables are **conjugate** to each other.

The computation that allows you to extract the $a_i$'s from derivatives of $\log Z$ generalizes to an expression for the expected value of any polynomial in the $a_i$'s. In this way, $Z$ contains a large amount of information about the statistics of our set of thermodynamic variables, and in particular all their variances and covariances. I encourage you to check that we have

$$\mathbb{E}[A_{i_1} \cdots A_{i_n}] = \frac{(-1)^n}{Z} \frac{\partial^n Z}{\partial \lambda_{i_1} \cdots \partial \lambda_{i_n}}$$

$$\mathrm{Cov}(A_{i_1}, A_{i_2}) = \frac{\partial^2}{\partial \lambda_{i_1} \partial \lambda_{i_2}} \log Z.$$

(In particular, since covariance matrices are positive definite, this means $\log Z$ is convex.) This is one of many ways in which the statistical-mechanical picture contains strictly more information than the thermodynamic one. The values of thermodynamic variables have been identified

in our new framework with the *expected values* of random variables, and the new framework also contains information about variances, covariances, and higher moments of these variables. This is not just an artifact of the formalism: the variances that arise from this formula consitute a bona fide quantitative prediction of statistical mechanics that can be (and has been) checked by experiment.

In addition, suppose that we have some "control parameters" $b_1, \ldots, b_n$, that is, variables in the third group above. If we vary both the $a_i$'s and the $b_j$'s we can write the corresponding change in $S$ as

$$\frac{1}{k} dS = \sum_{i=1}^{m} \lambda_i da_i + \sum_{j=1}^{n} \gamma_j db_j,$$

defining $\gamma_j = (1/k)(\partial S/\partial b_j)$ by analogy with the formula for $\lambda_i$, and we also refer to the $b_j$ and $\gamma_j$ variables as "conjugate." Now, let's imagine energy is one of the $A_i$'s, say $A_m$, so that $a_m = E$ and $\lambda_m = \beta$. We can then rewrite this formula to look more like one we saw earlier:

$$dE = \frac{1}{k\beta} dS - \sum_{i=1}^{m-1} \frac{\lambda_i}{\beta} da_i - \sum_{j=1}^{n} \frac{\gamma_j}{\beta} db_j.$$

We are therefore led to identify $\lambda_i/\beta$ and $\gamma_j/\beta$ with the generalized pressure corresponding to the $A_i$ or $b_j$ variable.

## 3.7 The Second Law in Statistical Mechanics

Finally, we should discuss how to extract a version of the second law of thermodynamics. It is a simple consequence of Liouville's theorem that running time forward cannot change the entropy, and this leads to a common stumbling block when learning this machinery: how are we supposed to get entropy to increase, as it sometimes does in thermodynamics? The story can't be as simple as just tracking $S$ for a probability distribution over time, but there is still a story to tell.

Suppose our system starts in an equilibrium distribution $\mu_0$. Now, we change the constraints such that the equilibrium values of our thermodynamic variables are different, meaning that $\mu_0$ is no longer an equilibrium distribution. Allow the system to equilibrate by running time forward until the expected values of our variables have settled down to the new equilibrium values with low variances. (Recall that the fact that this happens is one of our basic assumptions!) Call the resulting distribution $\mu_0'$. By Liouville's theorem, $S[\mu_0'] = S[\mu_0]$.

Finally, we may also consider the distribution which, among all distributions with the same expected values of the variables as $\mu_0'$, has the largest possible entropy. Call this distribution $\mu_1$. Since, of course, $\mu_0'$ is one of the distributions satisfying this constraint, we have $S[\mu_1] \geq S[\mu_0'] = S[\mu_0]$. Because our system has equilibrated, for the sake of future predictions we are free to *replace* $\mu_0'$ with the equilibrium distribution $\mu_1$; this is the sense in which entropy is higher at the end of this process. A useful picture is that $\mu_0'$ "knows" not only the new equilibrium values of the variables, but also the fact that we started out with different ones. Now that we have equilibrated, this history is irrelevant and we are free to forget about it.

It is also possible to tell a version of this story for a system in thermal contact with its environment wherein you only perform this "forgetting" operation on the environment, rather than the system, yielding a picture where the total entropy can increase even though the system is

not at equilibrium. I encourage you to work out that, if you assume that the environment's temperature never changes, this recovers the picture from the thermodynamics section involving decreasing free energy.

This is a simple example of a more general procedure called **coarse-graining**. Basically all models of non-equilibrium processes in statistical mechanics implement a more sophisticated version of this idea, continuously projecting the probability distribution representing the current state onto some smaller space of distributions, throwing out fine-grained information about the state that is (so the model asserts) irrelevant to predicting its future macroscopic behavior.

No matter how you do it, getting entropy to increase in our setup requires that you throw away information about the exact state over time, hopefully because that information was useless for further predictions about the future. (This should really be seen as a feature, not a bug: in the information-theoretic context, increasing entropy *means* losing information.) Because the microscopic physics is reversible, retaining every single detail about the distribution means that in principle the original distribution can be recovered, and so there's no way entropy could possibly increase.

From this perspective, the second law is tightly connected to the equilibration hypothesis: it is the assertion that the system will eventually reach a state where the values of the thermodynamic variables are the only information that is useful for making predictions, and that with very high probability the resulting values of those variables depend hardly at all on the initial state.

## 3.8 The Ideal Gas and the Gibbs Paradox

As an example, we'll show how to extract the ideal gas law using statistical mechanics. (Recall that in pure thermodynamics, it just has to be taken as a postulate.) We start by writing down the Hamiltonian. What makes a gas "ideal" is that the gas molecules don't interact with each other, but if the molecules themselves are big enough they might have some degrees of freedom (like rotation, for example) that contribute to the total energy. For simplicity, we'll restrict our analysis to a *monatomic* ideal gas, where this doesn't happen, meaning that

$$H = \sum_{i=1}^{3N} \frac{p_i^2}{2m},$$

where $m$ is the mass of one gas particle.

All three of the ways we listed earlier for a variable can be specified appear in this setting. The energy $E$ has a specified expected value, the number of particles $N$ is specified exactly, and the volume $V$ of the container will be treated as a control parameter. (We can imagine the volume appearing in the Hamiltonian as a big spike in the potential energy around the edges of the container, making any states with particles outside the container contribute an exponentially small amount to the partition function. We are simplifying the computation by using the above formula for $H$ and only integrating over states where all the particles are in the container.)

We will start by computing the partition function. We have:

$$\begin{aligned} Z(\beta) &= \int_X \exp(-\beta H(x)) \\ &= \int dq_1 \cdots dq_{3N} dp_1 \cdots dp_{3N} \exp\left(-\beta \sum_{i=1}^{3N} \frac{p_i^2}{2m}\right). \end{aligned}$$

Since the position coordinates don't appear in the integrand and the gas is confined to a container of volume $V$, each integral over the three spatial coordinates of a single particle just contributes a factor of $V$. The rest of the integral factors into $3N$ independent Gaussian integrals of the form $\int \exp(-\beta p^2/2m)dp$, and so

$$Z = V^N \left(\frac{2\pi m}{\beta}\right)^{3N/2} = V^N (2\pi mkT)^{3N/2}.$$

Our computations from before let us easily compute the energy and entropy:

$$E = -\frac{d}{d\beta}(\log Z) = \frac{3N}{2\beta} = \frac{3}{2}NkT;$$

$$S = k(\log Z + \beta E) = Nk \log V + \frac{3Nk}{2}\log(2\pi mkT) + \frac{3Nk}{2}.$$

There are many ways to extract pressure from these thermodynamic variables. One is

$$P = \frac{1}{k\beta}\left.\frac{\partial S}{\partial V}\right|_E = NkT/V,$$

and so we see that $PV = NkT$ as desired.

At the start of the last section we mentioned that we have only pinned down the thermodynamic entropy up to an additive constant. This has no effect on our computation of the ideal gas law, but it is possible to construct situations which will force us to choose the constant "correctly."

First, there is a small conceptual problem with the way we've defined the partition function: $Z$ is an integral over phase space of a dimensionless quantity, which means that it has units of phase space volume, and so it doesn't make sense to take its logarithm. We can solve this by dividing $Z$ by some constant with these same units, which will have the effect of subtracting the logarithm of that constant from $S$. Nothing about our setup so far forces any particular choice on us, but it's conventional to divide $Z$ by $h^{3N}$ (where $h = 2\pi\hbar$ is Planck's constant) in order to make the results agree with the predictions of quantum statistical mechanics in the high-temperature limit.

Much more serious is the problem of the dependence on $N$. Imagine two identical containers of the same ideal gas side by side separated by a removable wall. Say each container has $N$ particles, volume $V$ and entropy $S$. Because the two systems are independent from one another, the total entropy must be $2S$. Now, remove the wall and allow the combined system to come to equilibrium. By plugging in $2N$ and $2V$ into the formula for the entropy above, we see that the entropy is now more than $2S$, which is a big problem: if $N$ is large then with high probability, half of the particles are on each side of the combined box, so if we reinsert the wall our state is now the same as the state we started in, which means its entropy would have to drop back *down* to $2S$, violating the second law. This is known as the **Gibbs paradox**.

The resolution is quite simple. If we regard the final state as identical to the initial state (as we should), then that means the gas particles are **indistinguishable**, that is, interchanging two of them does not change the state. This, in turn, means that in our formula for $Z$ we have overcounted the states by a factor of $N!$. I encourage you to show that dividing $Z$ by $h^{3N}N!$ and using Stirling's formula to approximate $\log N!$ yields the **Sackur-Tetrode formula**

$$S = Nk\left[\log\left(\frac{V}{N}\left(\frac{2\pi mkT}{h^2}\right)^{\frac{3}{2}}\right) + \frac{5}{2}\right].$$

The situation would be different if the two initial containers held *different kinds* of gas. In this situation, some of the gas particles would be distinguishable from one another, and the final state would actually be different from the initial state, that is, it would be physically correct for the removal of the wall to increase the entropy. The additional entropy that arises in this way is called the **entropy of mixing**.

The computation of the thermodynamic properties of an ideal gas that we have gone through in this section barely scratches the surface of what the statistical-mechanical machine can do. One has to make some assumptions on the way to the expression for the canonical distribution, most notably that the *microcanonical* distribution indeed describes the distribution of states you will find if you look at random samples of isolated systems at equilibrium. But whatever you think of these assumptions, the fact is that the resulting theory leads to an astonishingly large number of very well-confirmed predictions. I hope to cover more of them in future articles in this series.